NEURIPS 2026

MAIN CONFERENCE
DEC 2026

OMNISE

EVERY ANSWER MUST SHOW ITS EVIDENCE!

FEATURING STR · RAC · AVET

SUPPORT-FAITHFUL OPTIMIZATION FOR EVIDENCE-CENTRIC AUDIO-VIDEO REASONING

Xilin He[1,2] · Yifan Shen[1] · Ankan Deria[1] · Xiangyu Yue[2,3] · Muhammad Haris Khan[1]

[1]MBZUAI [2]MMLab, The Chinese University of Hong Kong [3]Shenzhen Loop Area Institute · xilin.he@mbzuai.ac.ae

NEURAL INFORMATION PROCESSING SYSTEMS

Mohamed bin Zayed University of Artificial Intelligence

香港中文大學 The Chinese University of Hong Kong

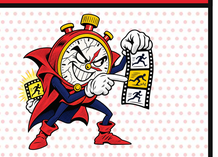
深圳河套學院 Shenzhen Loop Area Institute

MEANWHILE, IN AUDIO-VIDEO QA...

RIGHT ANSWER, WRONG EVIDENCE!

Answer-centric RL rewards the **final answer**, not the evidence behind it. Three villains slip through:

WRONG SPAN



looks at the wrong moment

SHORTCUT

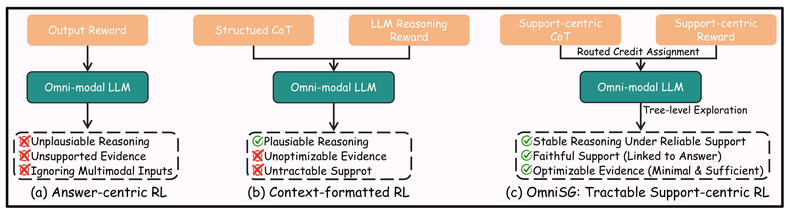


trusts only one modality

PHANTOM LINK



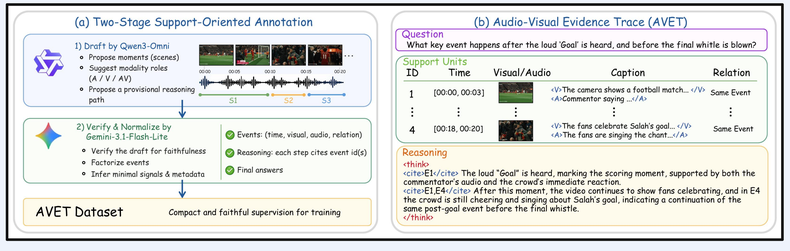
invents an audio-visual link



ORIGIN STORY

AVET: THE EVIDENCE TRACE

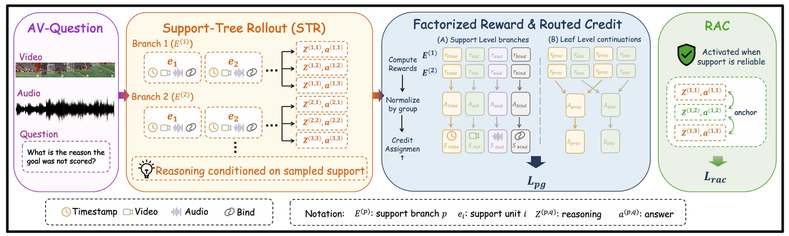
Each QA pair becomes a trace: **evidence units** (time · visual · audio · relation) + reasoning steps that **cite** them.



26K! AVET TRACES

traces seed the SFT cold start; RL then optimizes them directly.

THE TEAM ASSEMBLES: OMNISE



SUPPORT-TREE ROLLOUT — Explore evidence branches first, then reason on each one.

FACTORIZED REWARD — Score time, visual, audio, binding, reasoning and answer, each only on its own tokens.

RAC — On reliable evidence, sibling rollouts are pulled toward the best one.

SHOWDOWN ON DAILY-OMNI!

Question
Which audio events correspond to the indoor shooting range activities shown in the video?



Options:
A. Countdown sequence during target practice
B. Electronic music overlapping with gunshots
C. Hip-hop music with whoosh sounds
D. Wind sounds followed by human speech

HumanOmniv2

<context>The video shows police officers at an indoor shooting range. There is a countdown sequence during target practice, and the audio features a mix of electronic music and gunshots.</context>
<think>As I watch the video, I notice the police officers at an indoor shooting range. The scene is set for target practice, and I can hear a distinct countdown sequence, which is a common element in such activities to build tension and focus. Listening to the audio, I notice a mix of electronic music playing in the background, which is unusual but not uncommon in training environments to create a certain atmosphere. Additionally, the sound of gunshots punctuates the audio, confirming the shooting context. While considering the options, I observe that option A mentions a countdown sequence and gunshots, which aligns perfectly with what I see and hear. Options B, C, and D do not fit the scene as well because they mention different types of music or environmental sounds that are not prominent or relevant in this specific setting. Therefore, after analyzing the visual and audio cues, I conclude that the correct answer is A.</think>
<answer>A</answer>

OmniSE

<event id="E1"> <time>00:08-00:15</time> <visual>The person is holding a handgun and firing it towards the targets.</visual> <audio>A series of distinct, sharp gunshots are heard.</audio><bind>same_event</bind> </event> <event id="E2"> <time>00:25-00:29</time> <visual>The person continues to practice shooting at the range.</visual> <audio>Electronic music plays in the background.</audio><bind>offscreen_audio</bind></event>
<think><step id="R1" cite="E1">The visual of the person firing a handgun at a range corresponds directly to the audible gunshots.</step>
<step id="R2" cite="E2">The electronic music heard in the background while the person is shooting is consistent with the provided options.</step>
<step id="R3" cite="E1,E2">Combining the visual of the shooting activity with the audible gunshots and background music confirms that the audio events correspond to the shooting range activities.</step></think><answer>B. Electronic music overlapping with gunshots</answer>

PHANTOM! Baseline hears a countdown that isn't there: A ✗

NAILED IT! OmniSE cites the gunshots and the music: B ✓

THE FINAL BATTLE: RESULTS

Gain over our SFT starting point (7B)

+9.4! DAILY-OMNI

+4.8! WORLDSENSE

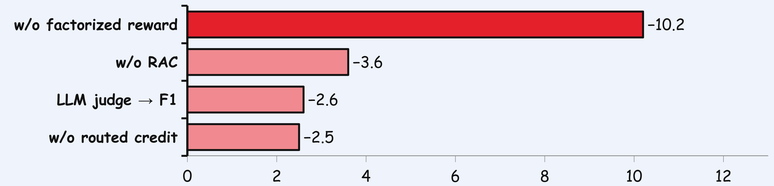
+15.0! VIDEOHOLMES

Accuracy (%)	Daily-Omni	WorldSense	IntentBench	VideoHolmes
*Qwen3-Omni-30B**	*63.6*	*54.0*	*65.7*	*49.0*
HumanOmniV2-7B	58.5	47.1	**69.3**	39.0
AVATAR-7B	55.7	46.0	61.9	**45.1**
AVET-SFT	55.6	47.8	65.1	23.2
Plain GRPO	55.1	46.5	62.8	23.4
OmniSE (ours)	**65.0**	**52.6**	**67.5**	**38.2**

**30B-A3B-Instruct reference. Bold: best 7B.*

WITHOUT THEIR POWERS...

Daily-Omni drop when a part is removed



Fewer hallucinations: AVHBench 72.16 → **74.86**

TO BE CONTINUED...

- ★ Optimize the **evidence**, not just the answer.
- ★ Each reward goes only to the tokens that earned it.
- ★ 7B OmniSE beats Qwen3-Omni-30B-A3B-Instruct on Daily-Omni: **65.0 vs 63.6**